Storage Technology

(with a data center bias)

Shigeki Misawa

Scientific Data and Computing Center

Brookhaven National Laboratory

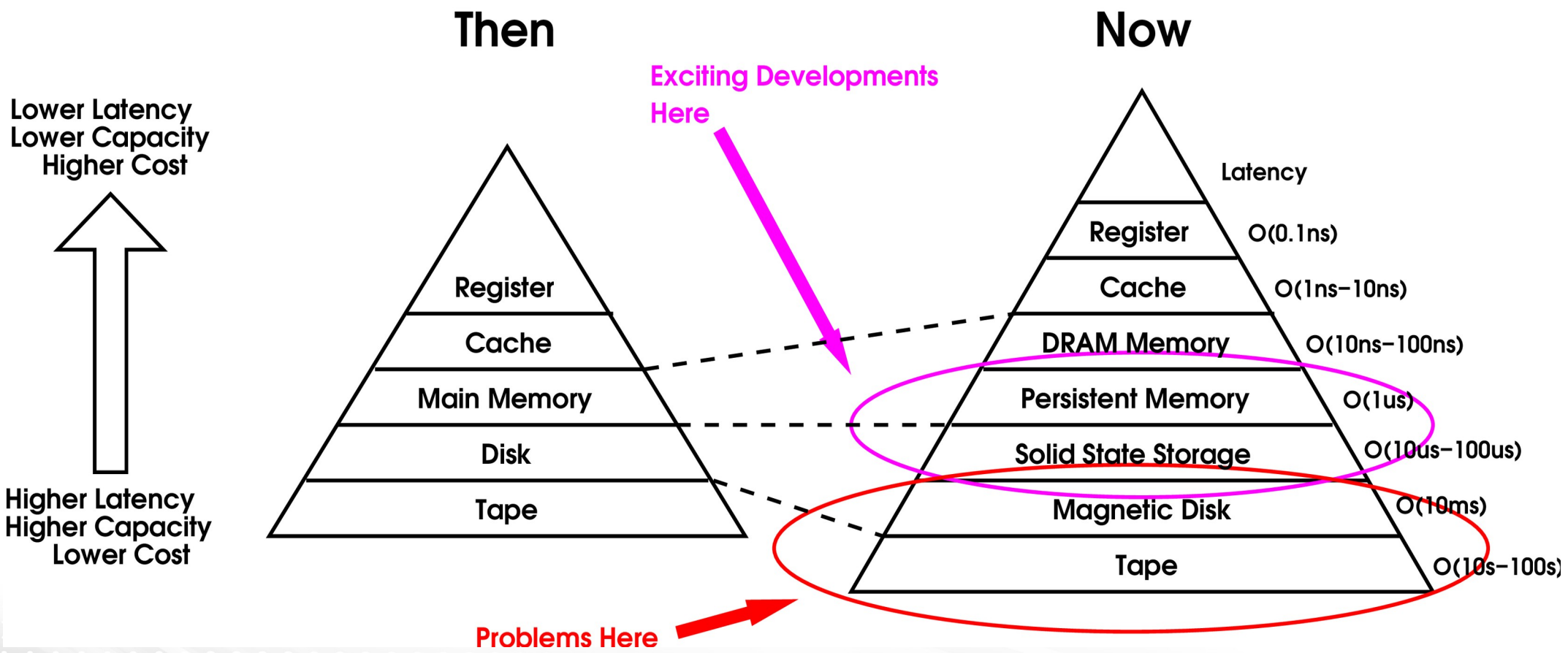
August 10, 2020

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

Storage Hierarchy: Changes in Progress



Persistent Memory

- Non volatile memory (NVM) on memory bus for ultra low access latency
  - 3D XPoint/low latency flash underlying NVM technologies
  - NVDIMM-N – NVM backed DRAM, byte access, looks like DRAM
  - NVDIMM-P – NVM w/ or w/o cache, byte access, JEDEC STD, access behavior slightly different from DRAM
  - DDR-T – Intel proprietary protocol for Optane DIMM
- Application re-write required for full benefit.
  - Multiple programming models
  - Can use as fast OS block device in the short term.

Solid State “Disk”

- Underlying 3D NAND Flash continues to advance
  - Cost/TB dropping, capacity increasing.
- Solid state disk (3D NAND and 3D XPoint)
  - Basis for high performance, low latency storage
  - NVM Express (NVMe) interface of choice
    - M.2/U.2/PCI-e mechanical interface
    - PCI-e Gen 4 devices starting to show up
  - EDSFF - New physical form factor
  - Cost/TB 5x to 10x higher than high capacity nearline disk

Solid State Storage Technologies

- NVMe Zoned Namespaces (ZNS)
  - Exposes page erase/write restriction of flash, currently hidden by the storage device flash translation layer (FTL), to OS/application
  - Additional visibility and control enables optimization for greater efficiency and improved performance
  - Requires software to manage tasks previously handled by the FTL
- NVMe over Fabric (NVMeOF)
  - NVMe over Ethernet/IB/FibreChannel/RDMA/TCP
  - Can be a basis for scale out NVMe storage systems and disaggregated computing systems.

Solid State Storage Systems

- SSD’s replacing HDD in most high IOPS environments
- Numerous high performance, scale out storage appliances are now available based on flash and 3D XPoint
  - Raw storage capacity typically lower than HDD systems due to cost, but most systems compensate with deduplication and compression.
- Expect continued development and deployment of solid state storage as costs drop, software and hardware interface specifications mature, and software is modified or written from scratch to utilize the capabilities of flash/3D XPoint.

Solid State Storage Systems and HEP

- SSDs deployed in high IOPS environments (e.g. databases, metadata servers, caches, and buffers)
- Unclear use cases for extremely high performance all flash systems in the data center for HEP, further handicapped by cost and capacity constraints. However, this may change over time.
- Utilization of PM and solid state storage in HEP face similar questions to GPU usage. Do the benefits of rewriting applications outweigh the software development costs or are sufficient benefits obtained without software modifications.

Magnetic Disks (HDD)

- Magnetic disks at a crossroad
  - Current recording technology (PMR/CMR) at areal density limit
  - Near or at limit of number of platters in 3.5” form factor
  - IOPS/TB continues to drop
    - (10 IOPS/TB for 8TB drive vs 5 IOPS/TB for 16 TB)
  - 3 HDD manufacturers left. Share by units shipped roughly 2:2:1
  - Manufacturers transitioning to high capacity “nearline” drives as SSDs displace HDDs in other markets. (Nearline HDD 55% of total revenues)
  - ~Half of all HDD, nearly all nearline, purchased by hyperscalers.

Future of HDD

- Transition to heat/microwave assisted recording (HAMR/MAMR) expected at some point (vendor dependent)
  - Enables higher areal bit density, and higher drive capacity
  - Product availability has been delayed several times over the years
- Multi-actuator (read/write heads) drives have been demonstrated
  - Increases IOPS/TB
  - Increases streaming read/write bandwidth per TB (BW/TB)
  - Single drive appears and performs like multiple independent disks
- Conflicting industry signals on cost/TB trends

Shingled HDD (SMR)

- Platter partitioned into zones. Overlapping data tracks within a zone
  - ~20% higher capacity compared to non SMR disks
  - Read performance identical to non SMR disks
  - Modifying data in a zone requires rewriting entire zone
- Can be applied to PMR, HAMR, and MAMR recording technologies
- SMR drive types
  - Host managed – Host must adhere to write restrictions.
  - Drive managed – HDD hides write restrictions from host
  - Host aware – Like drive managed, but can be host managed

Tape Technologies

- Linear Tape Open (LTO)
  - LTO-8 (Available now) – 12TB cartridge, 360MB/sec
  - LTO-9 (Expected late 2020) – 24TB native capacity
- IBM Enterprise Tape (TS-11XX)
  - TS1160 (Available now) – 20TB cartridge, 400MB/sec
  - Capabilities beyond LTO, e.g. faster seeks, higher capacity
  - Substantially smaller market share vs LTO
- Note that IBM is the sole supplier of leading edge tape drives

Tape Technology (cont'd)

- Clear technology roadmap with 400TB cartridges demonstrated in the lab.
  - LTO roadmap has LTO-12 native capacity at up to 192 TB
  - Historically, 2 to 3 year between tape generations
- Tape currently lowest cost/TB technology at large scale.
- SpectraLogic expects moderate improvements in read/write bandwidth (e.g. 500MB/sec projected for LTO-11)
- Tape bandwidth per TB is decrease (Tape drive performance not keeping up with tape cartridge capacity)

Limitations of Tape

- Tape is optimized for sequential access
  - Random reads and skipping files severely degrades read performance.
- When writing, data source must be able to keep up with tape drive for maximum performance.
- Physical layout of data is serpentine, requiring multiple end to end passes to read (or write) an entire tape. Logical layout is linear.
  - When sparse reading files, physically closest file to read may not be the logically closest.

Limitations of Tape and Tape Systems

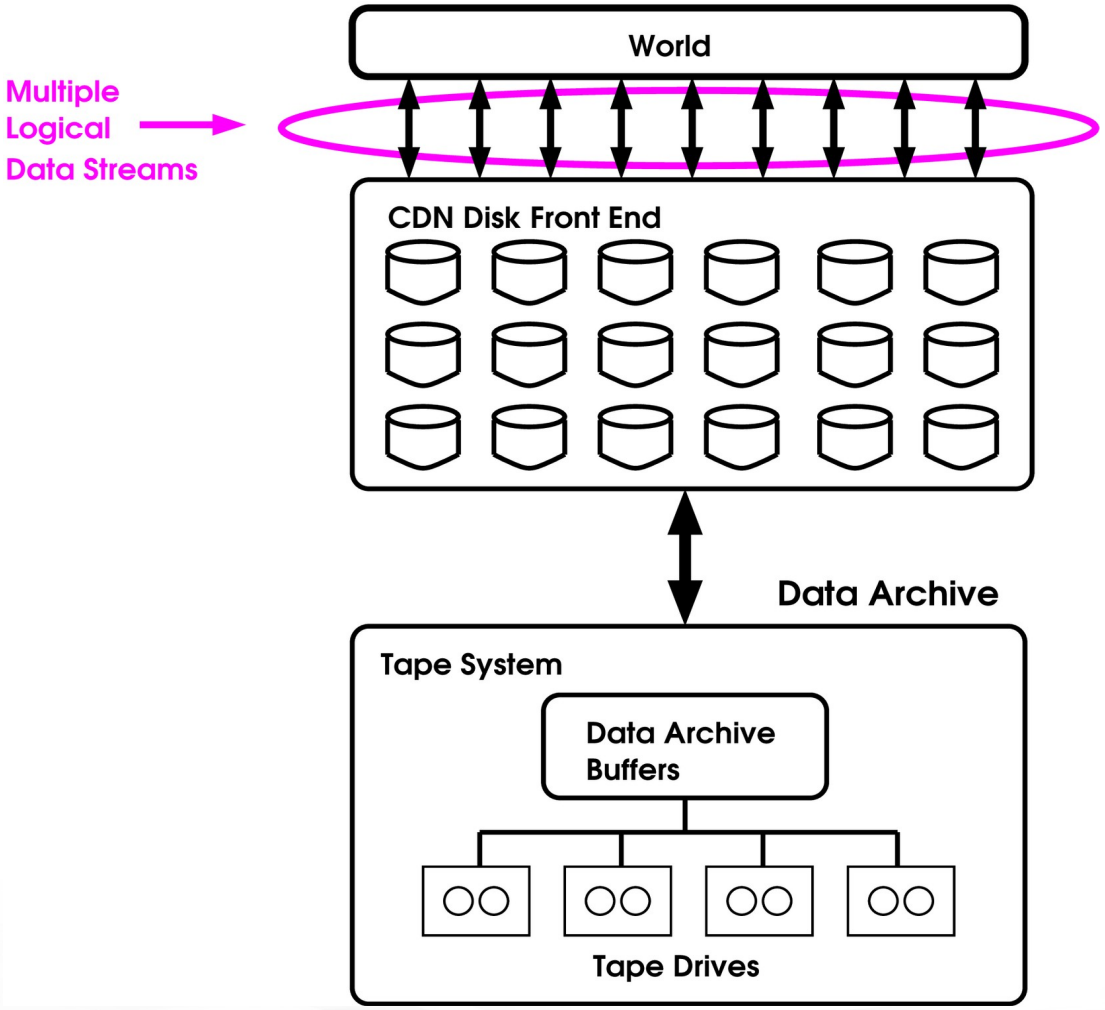
- For “simplistic” writes, tape drive needs to write data for at least 10 sec to achieve reasonable performance.
  - “Minimum” file size = Max write BW x 10 sec (~3GB for LTO-8)
- Capabilities of a tape system are highly dependent on the software used.
  - Can utilize advanced features of the tape drive, e.g., recommended access order (RAO), buffered tape marks
  - Can work around limitations of the tape, e.g., aggregation of small files

Storage and HEP

- HEP experiments are dependent on a worldwide content delivery network (CDN) to deliver data to compute resources
- On line storage (nearline disk or flash) sprinkled throughout the CDN with a handful of large data centers hosting 10’s of PB on line storage
- CDN front ends data archives hosted at a handful of large data centers
  - Each data archive host O(100PB) of data
  - Archives implemented with robotic tape libraries

Storage at a Major HEP Data Center

- Large HEP data centers are sending and receiving multiple streams of data at any given time
  - High capacity disk front end source/sinks the data
- Concurrently, the disk front end is reading/writing data from/to the data archive



CDN Disk Storage in HEP

- CDN disk systems are a critical component
  - Provide relatively low latency/high bandwidth access to data in contrast to tape
- Cost of CDN disk in the HL-LHC era is a problem
- ATLAS requires 7x more disk than budget allows in HL running
  - Assumes flat funding and 15% annual decrease in cost/TB
- Recent comments from Seagate suggest assumed cost/TB evolution may be too optimistic.

Data Archives for HEP

- Tape is the current technology of choice for HEP data archives
  - Lowest cost/TB
  - Theoretical media longevity (shelf life vs active life)
  - “Air gap” storage. Isolation from electrical and security threats
- Tape expected to compensate for short fall in CDN disk capacity
  - Likely requires higher read bandwidth per TB in the future
  - Effects on media longevity unclear

Future of CDN Disk Storage for HEP

- Evolution of hardware technology alone won't solve problems stemming from HL-LHC data volumes
- More advanced storage system software and more efficient utilization of storage resources are necessary, e.g,.
  - Tiered storage, HSM, erasure codes, QOS, advanced caching techniques, "deduplication"
  - Analysis "trains", "event service", "bulk" (efficient) data access, efficient data formats

Future of CDN Disk Storage for HEP

- Is there sufficient IOPS and BW for the HL-LHC era ?
  - With naive scaling IOPS and BW requirements scale with capacity requirements
  - 7x shortfall in capacity implies 7x increase in IOPS and BW required from disk that are purchased. Is there sufficient headroom ?
  - Effects of increased use of tape may also require additional IOPS and bandwidth.
- Influence of hyperscalers on technology evolution are unclear
  - Hyperscaler requirements may diverge from HEP requirements, e.g. Dropbox migration to SMR using bespoke software

Future of Tape for HEP

- Road map for tape capacity suggests storage capacity is not a technical problem. However, media cost is a concern
  - Market economics needs to be watched
- Bandwidth requirements a concern for HL-LHC
  - 10x increase in data volume naively translates to 10x increase in bandwidth requirements.
  - Expectation of LTO-11 bandwidth to be 1.5x to 3x higher than LTO-8
  - Implies 3x to 6x more LTO-11 tape drives to meet needs.
  - Doesn’t include additional drives needed handle more aggressive use of tape to compensate for disk shortfall

Future of Tape for HEP (cont'd)

- Elimination of sparse reading of tape, i.e., skipping unread files, is a high priority optimization for HEP storage stack.
  - Source of file skipping is interleaving of different types of data (logical data streams), received by the data center, on tape.
- If necessary, upgrade tape system software to utilize advanced tape drive features and where possible, work around limitations of tape.
- If necessary, optimize tape system software to operate tape drives at maximum performance.

Summary

- Like the growth in application specific processors in computing, there is a growing diversity of storage devices.
- Software, more than ever, are critical in the use of storage
  - “abstraction” layers limits achievable performance
  - Maximum capability is obtained from these new systems by tailoring software to the systems.
  - Additional performance is obtained by tailoring I/O to the capabilities of the devices.